Piecewise Linear Analysis of Biological Trajectories

Daniel Barton (SM09)

PhD supervisor: Dr. Jure Dobnikar
Institute of Physics, Chinese Academy of Sciences, Beijing

Twitching Motility

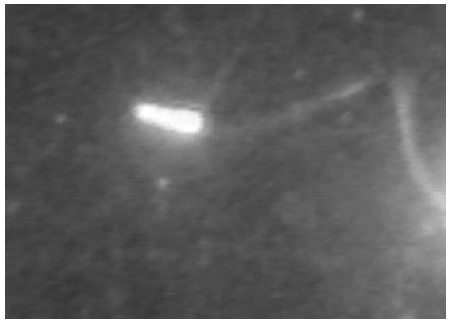


Fig 1. Skerker & Berg, 2001

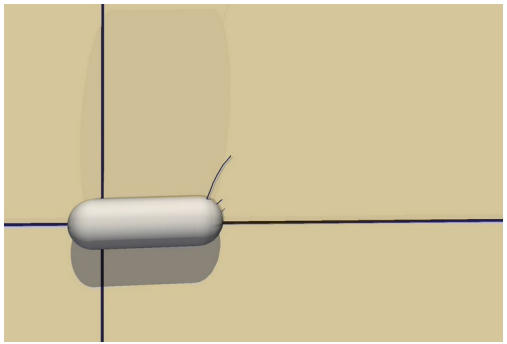


Fig 3. Simulation

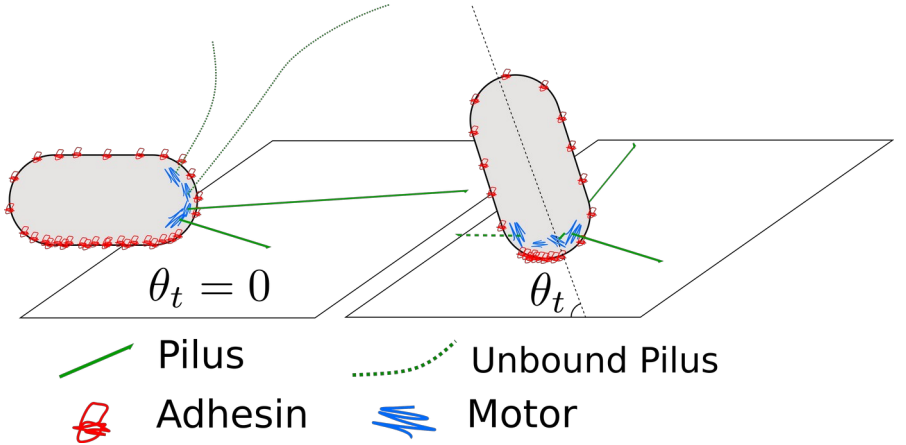


Fig 2. Simplified Drawing

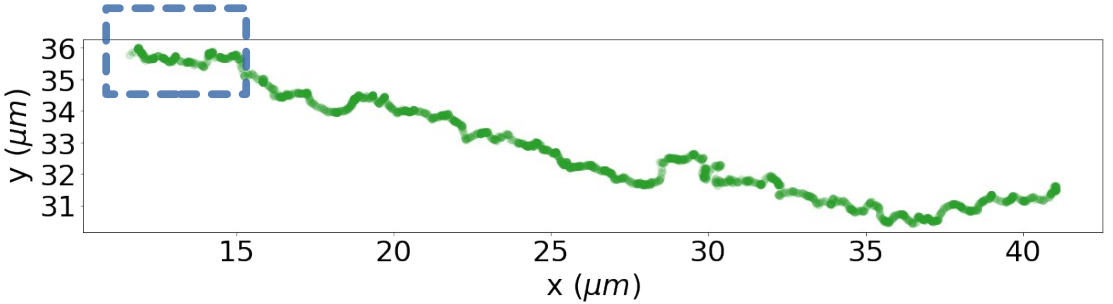


Fig 4. Experimental Trajectory

Tracking data

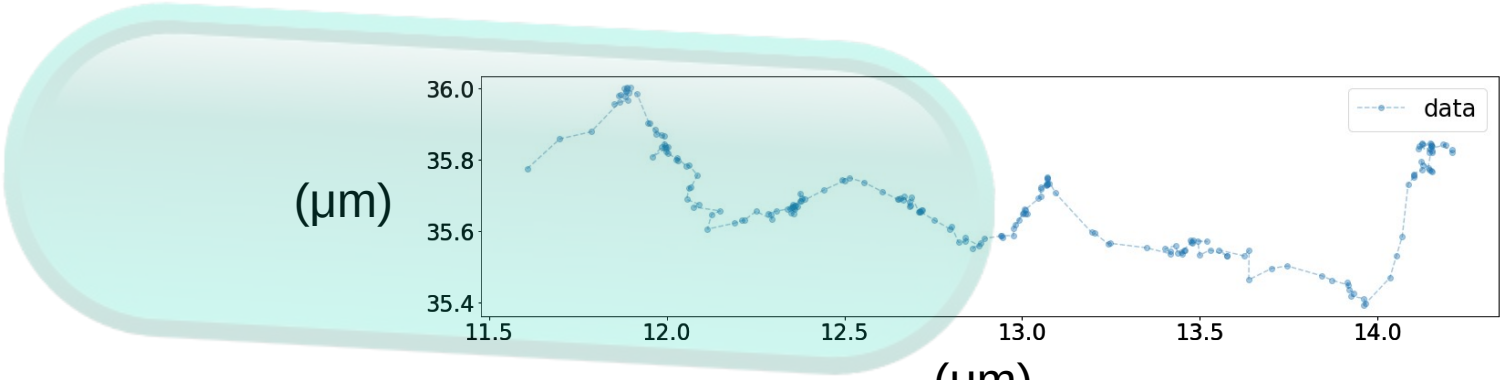


Fig 1. 20 seconds of Tracking Data

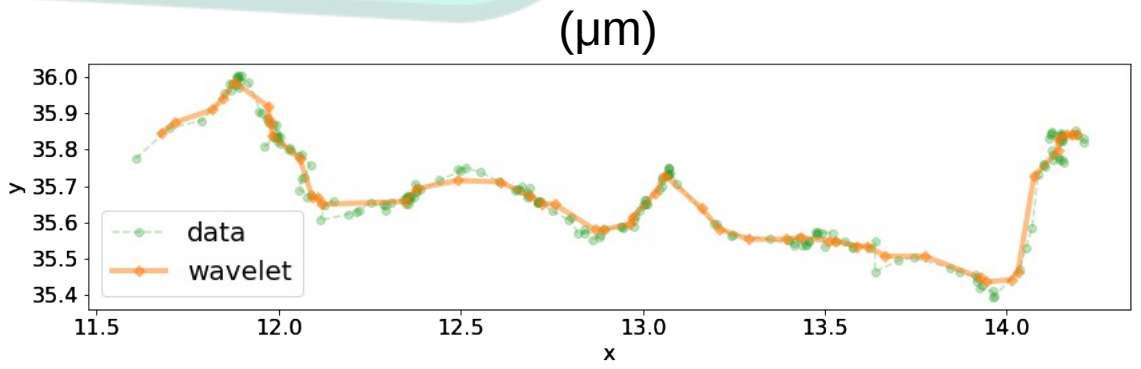
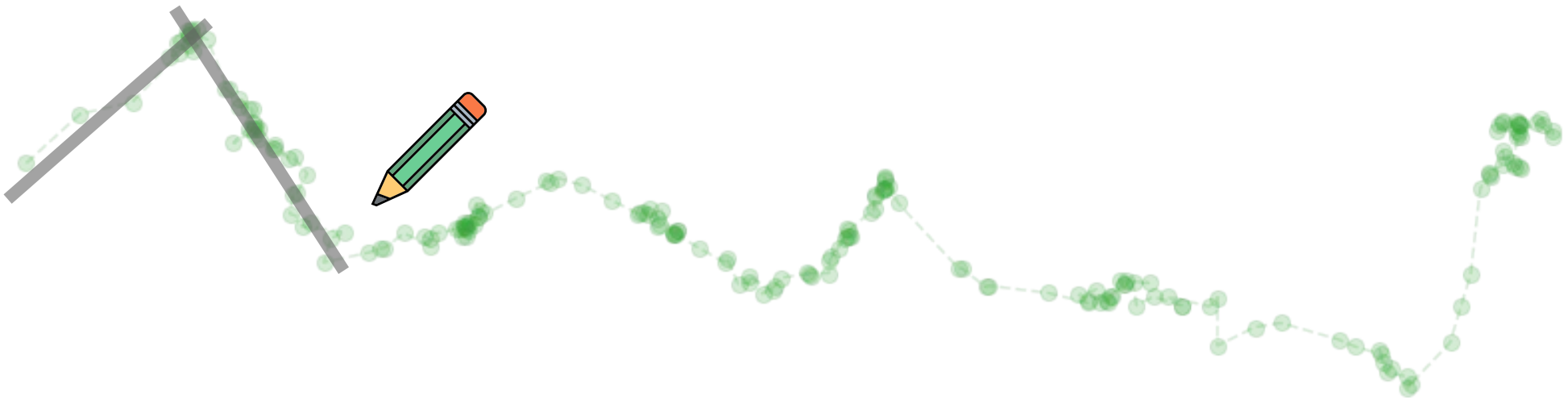


Fig 2. tracking data with wavelet smoothing

Question: Can we get more useful, more interpret-able information from this trajectory data?

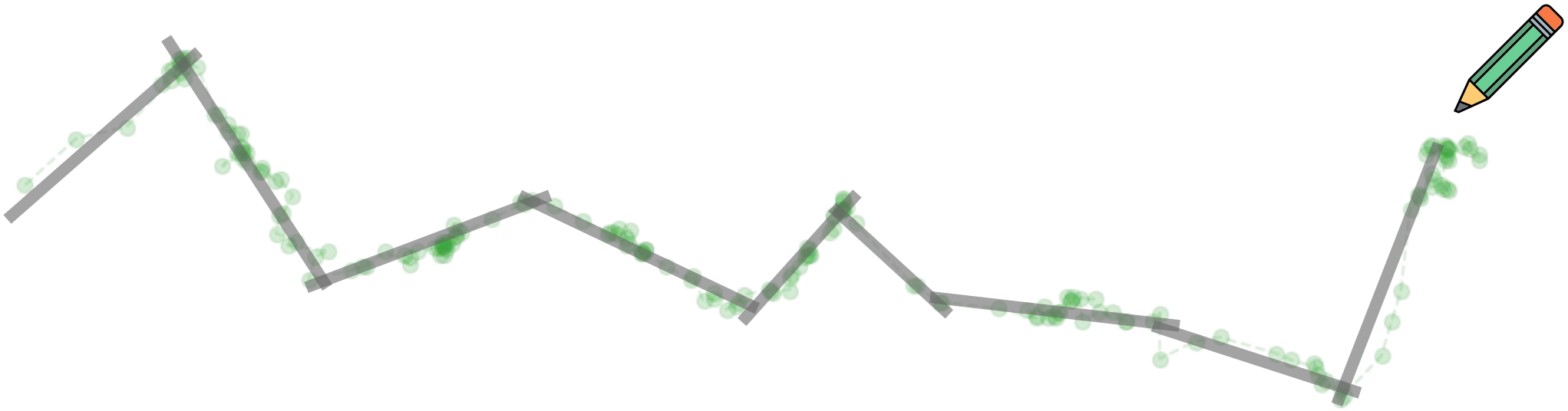
Linear Displacements

- Notice that the trajectory appears to be made up of roughly linear displacements.
- We imagine these displacements are connected to retraction of active filaments.
- Lets try to identify them by eye first.



Linear Displacements

- Notice that the trajectory appears to be made up of roughly linear displacements.
- We imagine these displacements are connected to retraction of active filaments.
- Lets try to identify them by eye first.



Piecewise Linear Solve

Consider N measurements taken at equal time intervals.

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$$

Measurements have some noise, e.g.

$$x_i = x_i' + \epsilon_i, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Define a *partition* of this data as a series of M indices

$$(k_1, \ldots, k_M),\ \mathrm{k}_1 < k_2 < \ldots < k_M$$
$$k_1 = 0, k_M = N$$

for any sub-sequence of the data we could fit a linear model

$$\boldsymbol{y} = m\boldsymbol{u} + \boldsymbol{c}$$

where distance to the line of a single data point is

$$d_i = ||\boldsymbol{c} + \boldsymbol{u} \cdot (\boldsymbol{x}_i - \boldsymbol{c})\boldsymbol{u} - \boldsymbol{x}_i||$$

and its common to obtain **u**, **c** by least squares optimisation. Minimise

$$\phi_2(k_1, k_2) = \sum_{i=k_1}^{k_2} d_i^2$$

Lets define a global cost function

$$\Phi(k_1, ..., k_M) = \sum_{k_i} \phi(k_i, k_{i+1})$$

Draw your attention to the following two issues.

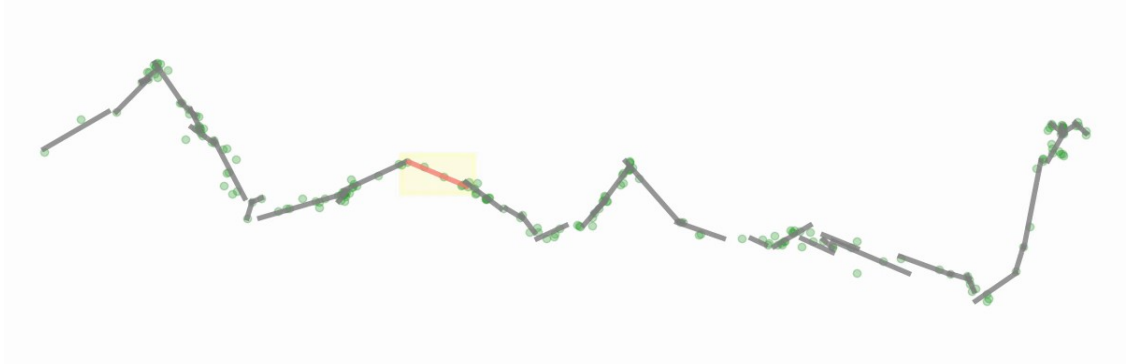
Issue 1: This cost function decreases with increasing M, in fact

$$M \to N, \Phi \to 0$$

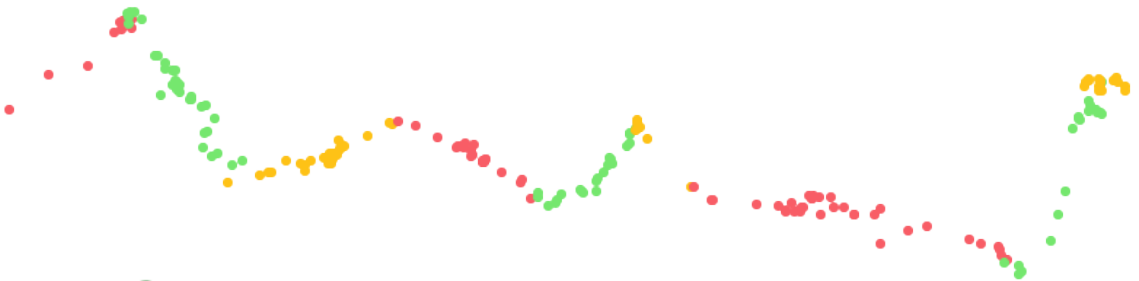
In regression and machine learning, this is called “overfitting”. We want M to be just large enough to capture the piecewise linear features of the data and no larger.

Issue 2: Global optimisation by searching all possible partitions becomes quickly infeasible for even moderate data size N.

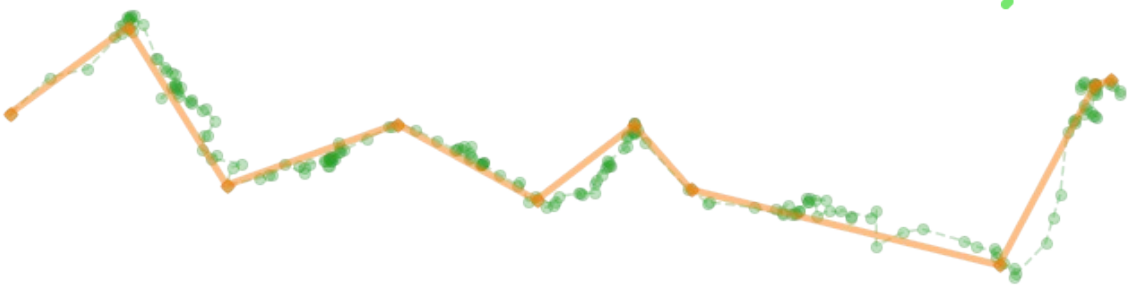
Piecewise Linear Solve



- Recursively join whole piecewise segments, choosing the join that has the minimum cost at each step.



- The result is a partition of the data into piecewise linear components.



- Connected model.

Minimum Description Length

lets define a new global cost function using an idea from information theory.

$$\text{let } \Phi_{\text{DL}} = M + \sum_i c_R(d_i), \quad c_R(d_i) = 1 \text{ if } d_i > R \text{ else } 0$$

The threshold parameter R is a bias/variance trade off, one idea is to select it from the inverse cumulative distribution function of the error distribution.

$$\mathcal{N}(0, \sigma^2),\ \sigma = 0.012 \qquad R = CDF^{-1}_{\mathcal{N}_\sigma}(0.99) = 0.028$$

Which says that there is a 1% chance that a measurement error will be greater than 0.028.

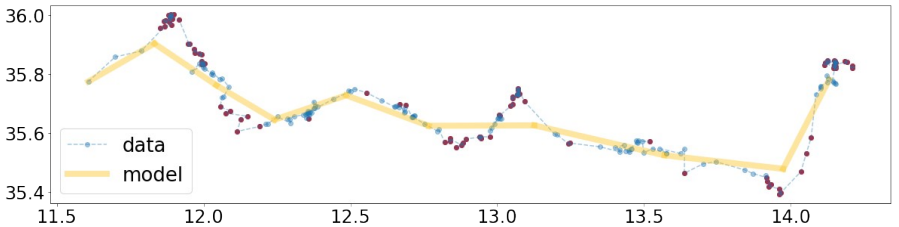


Fig 1. Candidate piecewise curve. Outliers are red and points close to the model are blue.

This time we used simulated annealing to stochastically search the space of all piecewise linear partitions of the data.

$$\text{Minimise } \Phi_{\text{DL}} \text{ for } \{M, k_1, ..., k_M\}, M \leq N$$

Unfortunately this method is quite expensive.

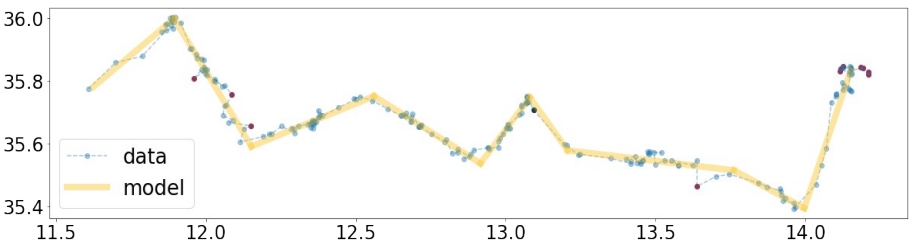
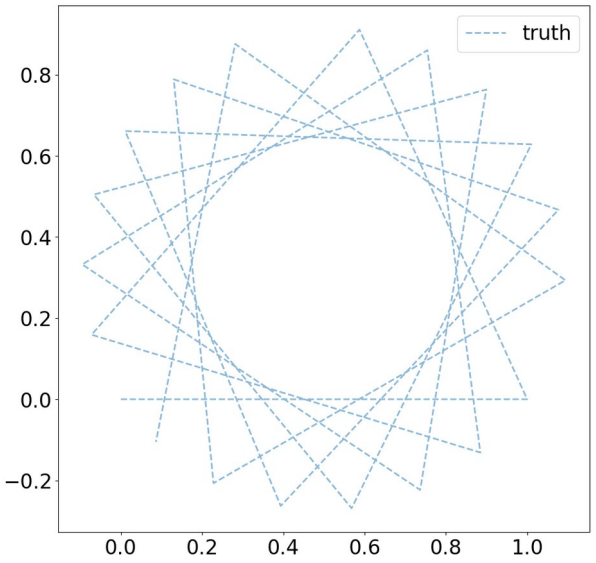
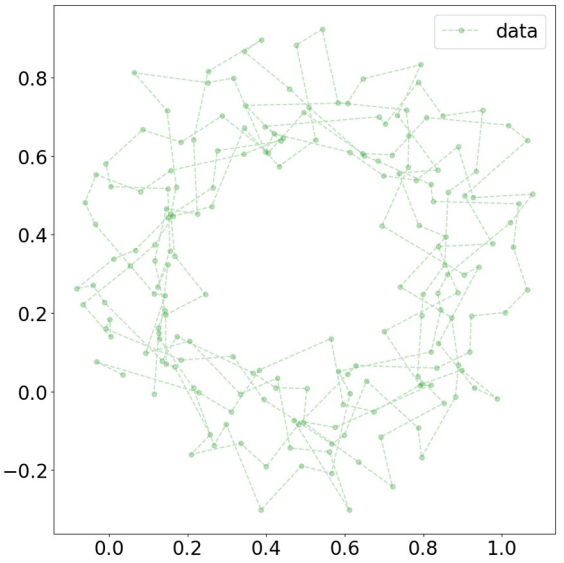
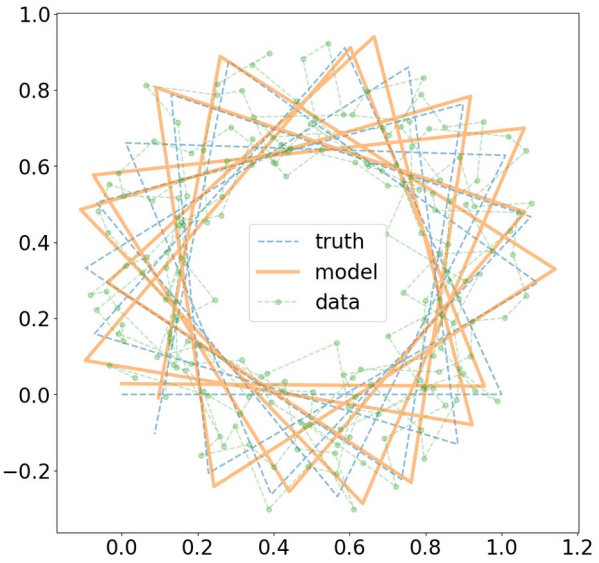


Fig 2. Minimum description length model.

Synthetic Data







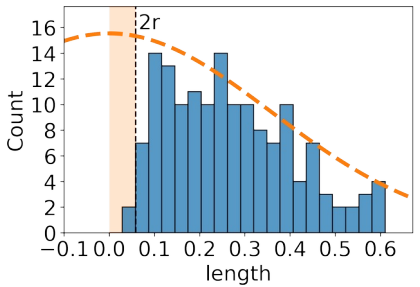
- Generate 10 random points per line segment.
- $\sigma = 0.1$

- Piecewise linear solution
- $R = 0.23$

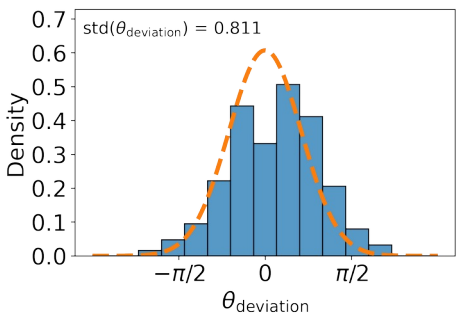
Statistics

This experimental trajectory has ~2000 data points

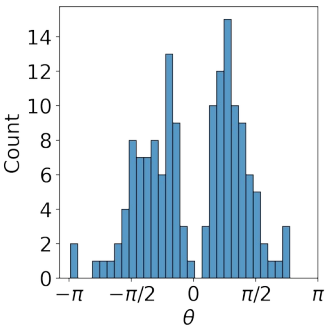
- Segment length distribution

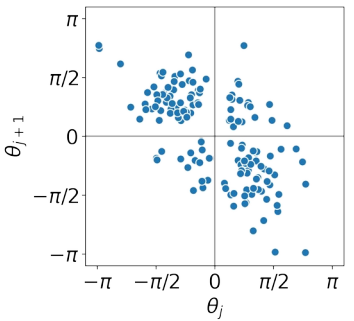


- Angle between body axis and velocity

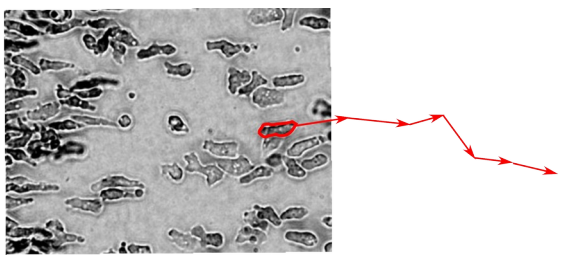


- Sequential linear displacements are anticorrelated (corrcoef = -0.65)





- This behaviour is shared with Dictyostelium (slime mold)



Summary

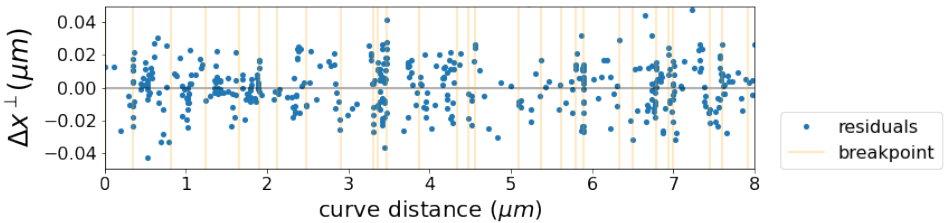
- Biological trajectories are often piecewise-continuous with high noise to signal ratio.
- Piecewise linear solvers exist for 1-dimensional problems, our algorithms are designed for N-dimensional trajectories.
- Extracting piecewise linear features of data (if they exist!) makes analysis much easier and more rewarding.
- This works for any trajectory data with piecewise structure!

Thanks for listening

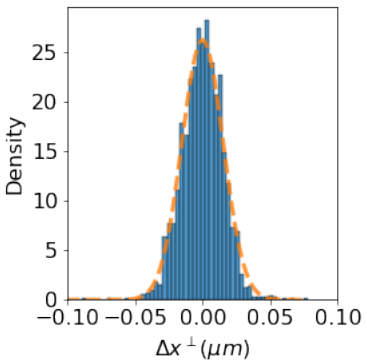
- Contact me: daluke.barton@gmail.com
- (SM09)

Residual Analysis

- Residuals vs. curve distance
- measurements clustering around breakpoints indicates low velocity/pausing between linear segments



- normal distribution of residuals



Estimated std : 0.0126

Residual std : 0.0152

“unexplained” deviation

0.0152 – 0.0126 = 0.0026

- Short timescale non-linearities would show up in the correlation between residuals.

$$\rho(\Delta x_i, \Delta x_{i+1}) = 0.285$$

$$\rho(\Delta x_i, \Delta x_{i+2}) = 0.087$$

- Very short timescale non linearities implies that the bacteria makes some displacements which are beyond our spatial and temporal resolution to detect.